DNA Sequence - The Journal of Sequencing and Mapping, Vol. 7(3/4), pp. 225-228
Reprints available directly from the publisher
Photocopying permitted by license only

Short Communication

Further Sequence Analysis of the DNA Regions with the *Rhodococcus* 20S Proteasome Structural Genes Reveals Extensive Homology with *Mycobacterium leprae*

ISTVÁN NAGY, SCHOOFS GEERT, VANDERLEYDEN JOS and RENÉ DE MOT*

F.A. Janssens Laboratory of Genetics, Catholic University of Leuven, B-3001 Heverlee, Belgium

Database Accession Nos.: Z82004, Z82005

(Received 10 July 1996)

The sequence of the respective DNA regions downstream of the 20S proteasome structural genes $prcB_1A_1$ (6 kb) and $prcB_2A_2$ (3.3 kb) of *Rhodococcus erythropolis* NI86/21 were determined. A highly conserved gene organization was observed between the two clusters which differed significantly in G + C content (68.8% versus 62.6%). Several ORFs were homologues of putative genes previously identified by genomic sequencing of the equivalent DNA in the related nocardioform actinomycete, *Mycobacterium leprae*, and thought to be specific for this pathogen. Three ORFs ($ORF8_1$, $ORF8_2$, $ORF12_1$) without a counterpart in *M. leprae* were found. No significant homology to known sequences including proteasome-related gene products was detected, except for $ORF9_1$ and $ORF9_2$ which display a high level of sequence identity with a partially sequenced ORF in *Streptomyces chrysomallus*. These downstream ORFs also show a significant level of sequence homology with the $ORF6_1$ and $ORF6_2$ which are located upstream of the proteasome structural genes in the respective clusters.

Keywords: Actinomycetes, gene organization, *Mycobacterium*, proteasome, *Rhodococcus*, *Streptomyces*

The eukaryotic 26S proteasome is the central multisubunit protease of the ubiquitin pathway of protein degradation and is formed by a 20S core complex and two polar 19S complexes (Hochstrasser, 1995; Goldberg *et al.*, 1995). The 20S proteasome constitutes the proteolytic core of this protease (Seemüller *et al.*, 1995) and is composed of fourteen related but different subunits, which are either of the α-type or of the β-type. The barrel-shaped 20S particle consists of four seven-membered rings with α-type subunits in the outer rings and β-type subunits in the inner rings (Lupas *et al.*, 1993). The 20S proteasome which was discovered in the archaebacterium *Thermoplasma acidophilum* is formed of only two subunits, α and β (Dahlmann *et al.*, 1989). The same architecture was reported for a second archaebacterial 20S proteasome, recently isolated from *Methanosarcina thermophila* (Maupin-Furlow

*Corresponding author. Tel.: 32-16-322398. Fax: 32-16-322990. E-mail: rene.demot@agr.kuleuven.ac.be.
Department of Applied Plant Sciences, Willem de Croylaan 42, B-3001 Heverlee, Belgium.

and Ferry, 1995). Remarkably, the first eubacterial 20S proteasome, discovered in *Rhodococcus erythropolis* NI86/21, contains two α and two β subunits (Tamura *et al.*, 1995). The α_1/β_1 and α_2/β_2 subunits are encoded by two unlinked gene pairs, $prcB_1A_1$ and $prcB_2A_2$, respectively. The subscript numbers refer to corresponding DNA regions containing the genes. Until now, no bacterial counterpart of the eukaryotic regulatory 19S particle or enzymes of the ubiquitin-conjugating machinery have been identified in bacteria, although there is evidence for the presence of ubiquitin in *T. acidophilum* (Wolf *et al.*, 1993) and *Anabaena variabilis* (Durner and Böger, 1995). As part of our search for such proteasome-related genes, we sequenced the DNA regions downstream of both proteasome gene pairs.

The 357 bp *Sma*I-*Bam*HI fragment of $prcA_1$ (Tamura *et al.*, 1995) was used as a probe in plaque hybridization to screen a λEMBL3 library of *Sau*3A-digested genomic DNA of *R. erythropolis* NI86/21. The insert of λFAJ2030 was found to contain about 6 kb of downstream sequence, the remainder of the insert overlapping with the previously characterized DNA fragment in λFAJ2028 (Nagy *et al.*, 1995; Tamura *et al.*, 1995). In addition, 3.3 kb of the region downstream of $prcA_2$, contained in λFAJ2029 (Tamura *et al.*, 1995), was sequenced. DNA sequencing of both strands on overlapping fragments subcloned in pUC19 was carried out with an automated sequencer (A.L.F., Pharmacia Biotech). The PC GENE software (IntelliGenetics) was used for sequence analyses. Potential coding regions were identified with the programs GCWIND (Shields *et al.*, 1992) and FRAME (Bibb *et al.*, 1984) Homology searches were performed using the FASTA, BLAST, and BLOCKS e-mail servers.

The gene organization in both clusters of *R. erythropolis* NI86/21 is shown in Figure 1. All ORFs, except $ORF8_1$ and $ORF8_2$, were located on the same strand as the structural proteasomal genes. $ORF10_1$, $ORF11_1$, and $ORF12_1$ may be translationally coupled since the stop and start codons of the adjacent ORFs overlap by two base-pairs. Apparently, the highly similar gene organization previously observed for the two proteasome gene clusters (Tamura *et al.*, 1995) extends into the downstream region. We previously pointed out that the two clusters differ significantly in GC content. The sequence data for the downstream regions confirm this observation. The DNA region with $prcB_2A_2$ (62.6%) has a significantly lower proportion of G and C than the $prcB_1A_1$ region (68.8%). The data from genomic sequencing of *Mycobacterium leprae* cosmid B2126 reveal a quite similar arrangement of genes and ORFs. Since the *M. leprae* genome has not yet been completely sequenced, and proteasomes have not yet been isolated from this pathogen, it is not known whether a second set of proteasome genes exists in *M. leprae* as well. Two major differences with *Rhodococcus* are apparent. No equivalent of $ORF8_1$ and $ORF8_2$ is found adjacent to C3_260 which represents the putative *prcA* gene of *M. leprae*. Instead, a 2.4-kb DNA region without obvious ORFs is present in *Mycobacterium*. Also, an equivalent of the rhodococcal $ORF12_1$ is missing. The GC content of the DNA region in *Mycobacterium* (59.2%) is closer to the value for the $prcB_2A_2$-containing fragment.

The extent of sequence conservation between individual ORFs is shown in Table I. The high level of sequence identity between *Rhodococcus* and *Mycobacterium* reflects the close phylogenetic relatedness of these bacteria, both belonging to the nocardioform actinomycete cluster of the high-GC gram-positive bacteria. It is likely that the preliminary sequence data for *M. leprae* contain a sequencing error between C2_219 and C3_265, and between C2_220 and C1_181 since the introduction of a frameshift in these parts extends the homology over the entire length with $ORF9_1$/$ORF9_2$ and $ORF11_1$, respectively. Remarkably, the downstream $ORF9_1$ and $ORF9_2$ also display significant homology (about 40% identity over the entire length) with the $ORF6_1$ and $ORF6_2$ (C1_172 in *Mycobacterium*) which are located upstream of the $prcB_1$ and $prcB_2$ genes, respectively (*prcB* in *Mycobacterium*). Apart from $ORF9_1$ and $ORF9_2$, no

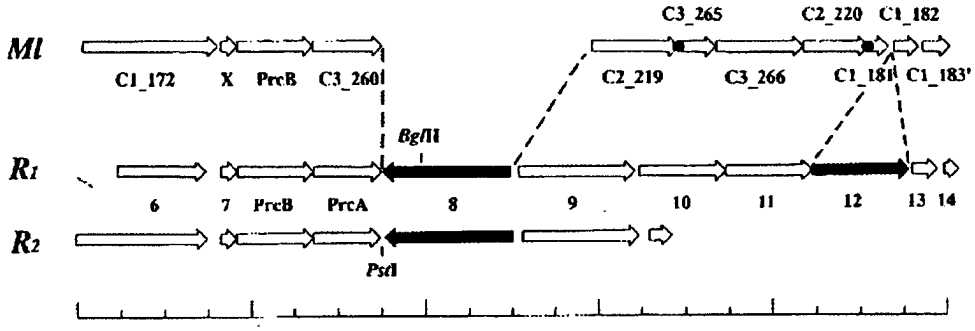

FIGURE 1 Gene organization in the DNA regions of *Rhodococcus erythropolis* NI86/21 containing the structural genes for the 20S proteasome: cluster R_1 with $prcB_1A_1$ and cluster R_2 with $prcB_2A_2$. The sequences of the DNA fragments carrying $prcB_1A_1$ (up to BglII site; accession number U26421) and $prcB_2A_2$ (up to PstI site; accession number U26422) were reported previously (Tamura *et al.*, 1995). The numbers refer to the different ORFs represented by arrows. The equivalent region of the *Mycobacterium leprae* (*Ml*) cosmid B2126 with its preliminary annotation (accession number U00017) is shown for comparison. Dashed lines are used to delineate blocks of gene organization conserved between *M. leprae* and *Rhodococcus*. The positions of putative framshifts in the *M. leprae* sequence are indicated with black dots. Filled arrows represent *Rhodococcus* ORFs with no counterpart in the equivalent *M. leprae* DNA region. One scale division represents 500 bp. The sequences have been submitted to the EMBL database with accession number *Z82004* and *Z82005*.

apparent homologues of the different rhodococcal and mycobacterial ORFs are currently known. The C-terminal parts of $ORF9_1$ and $ORF9_2$ display strong homology (64% identity in a 251 aa overlap) with the partially sequenced ORFA located upstream of the immunophilin gene *fkbB* of *Streptomyces chrysomallus* (Pahl and Keller, 1994). These independently obtained sequence data also suggest that C2_219 and C3_265 most probably form one contiguous ORF, as found for *Rhodococcus* for $ORF9_1$ and $ORF9_2$. Using the method of Dodd and Egan (1990) a potential helix-turn-helix motif (^{25}MSAAEAAAELGVTFTQLMSDLN^{46}) was predicted in the N-terminal part of $ORF11_1$, but no significant homology with known DNA binding proteins was found. Since for most ORFs in this region of the *M. leprae* DNA identified by genomic sequencing true homologues have now been iden-

TABLE I Homology between putative gene products from *Rhodococcus erythropolis* NI86/21 and *Mycobacterium leprae*. Clusters R_1 and R_2 refer to the DNA regions downstream of $prcA_1$ and $prcA_2$, respectively. N is used to indicate that no sequence data are available and A to denote the absence of a homologue.

Rhodococcus			*Mycobacterium*		
Cluster R_1	Cluster R_2	% Identity	Homologue	% Identity with homologue in	
				Cluster R_1	Cluster R_2
$ORF8_1$	$ORF8_2$	62.8% (487 aa)	A	—	—
$ORF9_1$	$ORF9_2$	95.1% (447 aa)	C2_219*	91% (302 aa)	90% (302 aa)
			C3_265†	84% (132aa)	85% (132 aa)
$ORF10_1$	$ORF10_2^1$	90.6% (85 aa)	C3_266	57% (330 aa)	72.9% (85 aa)
$ORF11_1$	N	—	C2_220*	62.6% (214 aa)	—
			C1_181†	43.8% (73 aa)	—
$ORF12_1$	N	—	A	—	—
$ORF13_1$	N	—	C1_182	50% (88 aa)	—
$ORF14_1^‡$	N	—	C1_183	75% (44 aa)	—

*Homologous to N-terminal part of *Rhodococcus* ORF(s)
†Homologus to C-terminal part of *Rhodococcus* ORF(s)
‡Truncated ORF (no further sequence data available)

tified in *R. erythropolis* NI86/21, a non-pathogenic nocardioform actinomycete, these ORFs should no longer be considered specific for this mycobacterial pathogen. However, at present, no predictions can be made about the possible functions of the characterized ORFs.

Acknowledgements

The authors thank D. Shields (Trinity College, Dublin) and G. Kleman (Ohio State University, USA) for providing the GCWIND and FRAME programs, respectively. U. Keller (Technische Universität Berlin, Germany) kindly communicated unpublished sequence data for *S. chrysomallus* ORFA. R.D.M. is a senior research associate with the Fund for Scientific Research (Vlaanderen).

Genomic sequencing for *Mycobacterium tuberculosis* (Accession No. Z73966) has revealed that the equivalent region is not linked to proteasome genes in this species.

References

Bibb, M. J., Findlay, P. R. and Johnson, M. W. (1984) "The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein coding sequences," *Gene*, 30, 157–166.

Dahlmann, B., Kopp, F., Kühn, L., Niedel, B., Pfeifer, G., Hegerl, R. and Baumeister, W. (1989) "The multicatalytic proteinase (prosome) is ubiquitous from eukaryotes to archaebacteria," *FEBS Lett.*, 251, 125–131.

Dodd, I. B. and Egan, J. B. (1990) "Improved detection of helix-turn-helix DNA-binding motifs in protein sequences," *Nucl. Acids Res.*, 18, 5019–5026.

Durner, J. and Böger, P. (1995) "Ubiquitin in the prokaryotic *Anabaena variabilis*," *J. Biol. Chem.*, 270, 28617–28622.

Goldberg, A. L., Stein, R. and Adams, J. (1995) "New insights into proteasome function: from archaebacteria to drug development," *Chemistry & Biology*, 2, 503–508.

Hochstrasser, M. (1995) "Ubiquitin, proteasomes, and the regulation of intracellular protein degradation," *Curr. Opinion Cell Biol.*, 7, 215–223.

Lupas, A., Koster, A. J. and Baumeister, W. (1993) "Structural features of 26S and 20S proteasomes," *Enzyme Protein*, 47, 252–273.

Maupin-Furlow, J. A. and Ferry, J. G. (1995) "A proteasome from the methanogenic archeon *Methanosarcina thermophila*," *J. Biol. Chem.*, 270, 28617–28622.

Nagy, I., Schoofs, G., Compernolle, F., Proost, P., Vanderleyden, J. and De Mot, R. (1995) "Degradation of the thiocarbamate herbicide EPTC (S-ethyl dipropylcarbomothioate) an biosafening by *Rhodococcus* sp. NI86/21 involve an inducible cytochrome P-450 system and aldehyde dehydrogenase," *J. Bacteriol.*, 177, 676–687.

Pahl, A. and Keller, U. (1994) "*Streptomyces chrysomallus* FKBP-33 is a novel immunophilin consisting of two FK506 binding domains; its gene is transcriptionally coupled to the FKBP-12 gene," *EMBO J.*, 13, 3472–3480.

Seemüller, E., Lupas, A., Stock, D., Löwe, J., Huber, R. and Baumeister, W. (1995) "Proteasome from *Thermoplasma acidophilum*: a threonine protease," *Science*, 268, 579–582.

Shields, D. C., Higgins, D. G. and Sharp, P. M. (1992) "GCWIND: a microcomputer program for identifying open reading frames according to codon positional G + C content," *Comput. Appl. Biosci.*, 8, 521–523.

Tamura, T., Nagy, I., Lupas, A., Lottspeich, F., Cejka, Z., Schoofs, G., Tanaka, K., De Mot, R. and Baumeister, W. (1995) "The first characterization of a eubacterial proteasome: the 20S complex of *Rhodococcus*," *Curr. Biol.*, 5, 766–774.

Wolf, S., Lottspeich, F. and Baumeister, W. (1993) "Ubiquitin found in the archaebacterium *Thermoplasma acidophilum*," *FEBS Lett.*, 326, 42–44.